Evaluating Regression Algorithms for Automatic Text Summarization in Brazilian Portuguese

Lucas Camargo Sodré[1], Hilário Tomaz Alves de Oliveira[2]

[1]Centro Universitário de João Pessoa – Unipê
João Pessoa – PB – Brazil

[2]Instituto Federal do Espírito Santo
Serra – ES – Brazil

{lucassodrebam8,hilariotomaz}@gmail.com

Abstract. *Automatic Text Summarization (ATS) is a prominent research area, which aims to automatically create a summary containing the most relevant information from one or more documents. One of the main challenges of ATS is to identify the most relevant information that should be included in the summary to be generated. This paper aims to analyze the application of regression algorithms to estimate the sentence relevance score of a collection of news articles written in Brazilian Portuguese in the ATS task. Experiments were performed to evaluate different sentence scoring methods, regression algorithms, and compare the results obtained with other works in the literature. The experimental results showed that the Bayesian regression algorithm obtained the best results based on the ROUGE evaluation measures, reaching a coverage rate of 62.09%.*

Resumo. *A Sumarização Automática de Textos (SAT) é uma proeminente área de pesquisa, cujo objetivo é a criação automática de um resumo contendo as informações mais relevantes a partir de um ou mais documentos. Um dos principais desafios da SAT é identificar as informações mais relevantes que devem ser inseridas no resumo a ser gerado. Este trabalho tem por objetivo analisar a aplicação de algoritmos de regressão para estimar um escore de relevância das frases de uma coleção de artigos de notícias escritos em português do Brasil na tarefa de sumarização. Experimentos foram executados para avaliar diferentes métodos de estimação de relevância das frases, algoritmos de regressão, e comparar os resultados obtidos com outros trabalhos da literatura. Os resultados experimentais demonstraram que o algoritmo de regressão Bayesiana obteve os melhores resultados com base nas medidas de avaliação do ROUGE, atingindo uma taxa de 62,09% de cobertura.*

1. Introdução

Com o avanço e popularização da tecnologia, tem ocorrido um aumento na criação e compartilhamento de informações digitais, nos mais diversos formatos, como textos, áudios, imagens e vídeos. Com tanta informação disponível, especialmente na Web, e o pouco tempo para analisá-las, as pessoas acabam tendo dificuldade para encontrar informações de interesse de maneira eficiente em certos cenários. Mesmo com o auxílio dos motores de busca na Web, como o Google[1], em diversas situações é necessário que os usuários leiam parcialmente ou integralmente os documentos retornados de uma consulta realizada.

[1]http://www.google.com

Um importante recurso usado pelos seres humanos para auxiliar na identificação de informações de interesse, especialmente em documentos textuais, são os resumos. Segundo o dicionário Michaelis[2], um resumo pode ser definido como "a apresentação sucinta de fatos, acontecimentos, entre outros, em que são abordados apenas os pontos principais, com o objetivo de transmitir uma ideia global de cada tópico tratado". Lendo um resumo é possível detectar as informações mais relevantes de um documento sem a necessidade de ler o seu conteúdo na integra.

Com o objetivo de tentar criar resumos automaticamente surgiu a Sumarização Automática de Textos (SAT) [Nenkova and McKeown 2012]. A SAT pode ser definida como a tarefa de criação automática de um resumo contendo as informações mais relevantes, a partir de um único documento (monodocumento) ou de um grupo de documentos relacionados (multidocumento) [Nenkova and McKeown 2012]. Os sistemas de sumarização podem ser classificados em duas grandes abordagens: Extrativa e Abstrativa. Os sistemas que seguem a abordagem extrativa identificam e selecionam as frases mais relevantes de um ou mais documentos, e as utilizam sem nenhuma alteração para a criação do resumo. Os sistemas abstrativos tentam simular a forma com que os seres humanos produzem resumos. Esses sistemas focam na seleção das informações mais relevantes dos documentos, e como expressá-las de uma nova forma.

Diante da complexidade no desenvolvimento de sistemas de SAT abstrativos, a sumarização extrativa ainda demanda muita pesquisa. O processo de SAT extrativo é, em geral, realizado em três etapas principais [Nenkova and McKeown 2012]: (i) Cria-se uma representação intermediária do(s) documentos usualmente aplicando técnicas de Processamento de Linguagem Natural (PLN); (ii) Computa-se a importância de elementos textuais como, por exemplo, n-gramas ou frases; e (iii) Geração do resumo através da seleção das frases mais relevantes dos documentos. Dois aspectos essenciais que precisam ser definidos nesse tipo de abordagem são: (i) Como mensurar a relevância dos elementos textuais; e (ii) Como evitar a inclusão de informações duplicadas no resumo gerado.

Diversos indicadores de relevância [Leite and Rino 2008, Oliveira et al. 2016] têm sido explorados na literatura para a tarefa de sumarização. Em geral, esses indicadores são baseados em técnicas estatísticas, como frequência, centralidade, ou em heurísticas, como posição das frases nos documentos. O trabalho desenvolvido por [Oliveira et al. 2016] avaliou diversas técnicas de ponderação da relevância das frases nas tarefas de SAT de artigos de notícias escritos em inglês. Os autores avaliaram os métodos individualmente e combinados, além de utilizá-los como atributos adotando algoritmos de classificação. Para documentos escritos em português, o trabalho de [Leite and Rino 2008] analisou diferentes métodos baseados em redes complexas também em conjunto com algoritmos de classificação.

Neste artigo, exploramos a estratégia de combinar esses indicadores usando algoritmos de regressão para estimar um escore de pontuação de importância das frases. O objetivo deste trabalho é analisar a aplicação de diferentes algoritmos de regressão para estimar a relevância das frases de uma coleção de documentos, visando selecionar aquelas que são mais relevantes para compor o resumo. Para isso, este trabalho é centrado na tarefa de sumarização extrativa de artigos de notícias escritos no Português do Brasil.

[2]http://michaelis.uol.com.br/moderno-portugues/

2. Sistema desenvolvido

O sistema desenvolvido é composto por três etapas principais, conforme ilustrado na Figura 1. Dada uma coleção de artigos de notícias em formato textual, primeiro realiza-se o pré-processamento dos documentos para estruturá-los em um formato adequado. Posteriormente, as técnicas de ponderação são aplicadas para mensurar a relevância das frases dos documentos de entrada. As técnicas adotadas são utilizadas em conjunto com um modelo de regressão previamente treinado para estimar um escore de relevância de uma dada frase. Por fim, as sentenças com maior pontuação de relevância são usadas para compor o resumo a ser gerado. Nesta etapa são aplicadas estratégias para evitar a inclusão de informações duplicadas e para ordenação das sentenças no resumo.

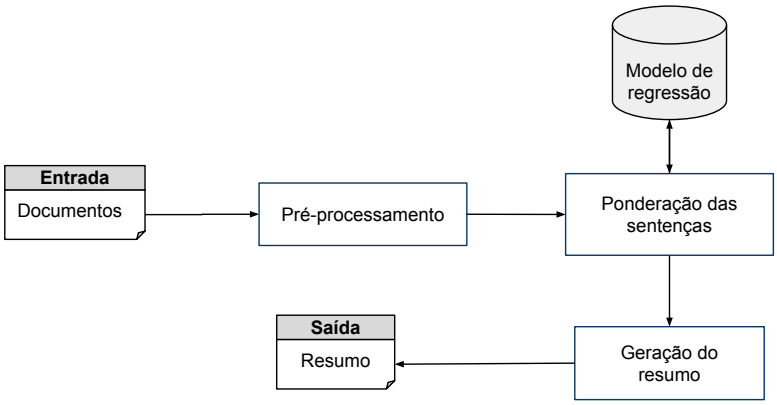

Figura 1. Etapas do processo de sumarização.

Na etapa de pré-processamento é realizada a estruturação dos documentos de entrada usando técnicas de PLN. Neste trabalho adotou-se a ferramenta Spacy[3] para execução das tarefas de segmentação das frases, tokenização, etiquetagem das classes gramaticais, reconhecimento de entidade nomeadas e remoção das *stopwords*. Outra tarefa de PLN usada foi a de *stemming*, que não é suportada pelo Spacy. Por isso, essa tarefa foi executada usando a ferramenta NLTK[4].

Na etapa de computação da relevância das sentenças, cada frase dos documentos de entrada é analisada e são computados onze indicadores de relevância de conteúdo que consideram aspectos, como frequência, centralidade, posição, entre outros. As técnicas utilizadas serão descritas na subseção a seguir. Essas técnicas são aplicadas em cada sentença da coleção de documentos e, posteriormente, aplica-se um modelo de regressão para estimar um escore de relevância para cada frase.

Posteriormente, é realizado o processo de geração do resumo. Para isso, as frases com maiores escores de relevância são selecionadas para compor o resumo. Para evitar redundância entre as sentenças selecionadas, utilizamos a heurística de que uma nova frase só é inserida no resumo caso ela não possua similaridade do cosseno maior que 0,5 com cada uma das sentenças já incluídas no resumo [Oliveira et al. 2016]. O processo de seleção é realizado de forma iterativa inserindo a frase com maior pontuação, desde que

[3]http://Spacy.io
[4]https://www.nltk.org/

ela respeite o critério de similaridade anterior, até que o tamanho máximo do resumo seja alcançado.

Além da escolha de quais frases serão selecionadas para compor o resumo é necessário definir a ordem em que elas serão inseridas. Para isso, as frases são ordenadas com base na sua aparição no documento ao qual ela pertence usando um índice que é atribuído a cada frase durante a etapa de pré-processamento. Contudo, duas frases com o mesmo índice podem ser selecionadas, por exemplo, a segunda sentença de dois documentos diferentes. Para resolver este conflito, foi utilizado o índice do documento ao qual a sentença pertence.

2.1. Métodos de ponderação da relevância das frases adotados

Nesta subseção são apresentadas as noves técnicas de ponderação das sentenças utilizadas neste trabalho. Essas técnicas foram selecionadas por terem apresentado bons resultados na literatura [Oliveira et al. 2016].

Frequência das Palavras. Esta é a técnica mais antiga aplicada para mensurar a relevância de uma frase para a tarefa de SAT [Oliveira et al. 2016]. A ideia deste método é que frases relevantes possuem palavras que são muito frequentes nos documentos a serem resumidos. Para isso, este método computa a quantidade de vezes que uma palavra aparece nos documentos. Neste trabalho adotamos a estratégia de não considerar *stopwords* e aplicamos a técnica de *stemming* antes de computar a frequência das palavras. A Equação 1 apresenta como o escore de relevância usando este método é calculado.

$$FreqPalavra(s_i) = \sum_{p_j \in P} frequencia(p_j) \tag{1}$$

- $frequencia$ retorna à frequência de uma palavra p_j pertencente a sentença s_i;
- P é conjunto de palavras que não são *stopwords* da frase s_i.

Frequência do Termo - Frequência Inversa das Sentenças (FT-FIS). Este método é uma adaptação do tradicional método *Term Frequency - Inverse Document Frequency* (TF-IDF) usado na área de Recuperação da Informação (RI) [Oliveira et al. 2016]. O método FT-FIS é computado como apresentado na Equação 2, já o escore de uma frase com base neste método é calculado como demonstrado na Equação 3.

$$FT - FIS(t_i) = frequencia(t_i) \times log(\frac{S}{Oc_{t_i}}) \tag{2}$$

$$FT - FIS(s_i) = \sum_{p_j \in P} FT - FIS(p_j) \tag{3}$$

- S é o total de sentenças da coleção de documentos;
- Oc_{t_i} é o total de frases que possuem o termo t_i.

Centralidade das sentenças. Este método pode ser definido como o grau de sobreposição entre uma frase em comparação com as outras sentenças dos documentos [Oliveira et al. 2016]. Esse método é baseado na hipótese de que frases que compartilham muitas informações com outras sentenças contêm informações relevantes. A Equação

4 demonstra como é computado este método. Além dessa formulação, adotamos uma variação deste método usando a medida de similaridade do cosseno de uma frase em relação as demais frases do conjunto de documentos. A similaridade do cosseno foi calculada comparando o vetor de palavras das frases, com *stopwords* sendo removidas e as palavras sendo representadas por seus radicais obtidos pelo processo de *stemming*.

$$Centralidade(s_i) = \frac{P_{s_i} \cap P_{s_c}}{P_{s_i} \cup P_{s_c}} \tag{4}$$

- P_{s_i} é o conjunto de palavras que não são *stopwords* da sentença s_i;
- P_{s_c} é o conjunto de palavras que não são *stopwords* das outras frases $s_i \neq s_c$ do(s) documento(s).

Entidades Nomeadas. Entidades nomeadas, geralmente, se refere a nomes de pessoas, lugares, organizações, entre outros [Oliveira et al. 2016]. A ideia deste método é que o resumo deve conter a maior quantidade possível de entidades nomeadas mencionadas no(s) documento(s). Para isso, este método atribui uma maior escore para sentenças que possuem mais entidades nomeadas. A Equação 5 demonstra como é computado o escore de uma frase.

$$EntidadesNomeadas(s_i) = \frac{\#total_de_entidades_em_s_i}{\#maximo_de_entidades_em_uma_sentenca} \tag{5}$$

Posição das sentenças. A posição das frases é um dos métodos que apresenta melhores resultados na SAT, principalmente em resumos de artigos de notícias [Oliveira et al. 2016]. Este método é baseado na ideia que as frases mais próximas do início do documento são mais importantes. Este método é computado conforme demonstrado na Equação 6.

$$Posicao(s_i) = 1 - \frac{i}{S_d} \tag{6}$$

- i é índice da posição da sentença s_j no documento d, com i começando em 0;
- S_d é o total de frases do documento d.

Similaridade com o título. Geralmente, o título de um documento apresenta indícios do tema abordado, principalmente em artigos de notícias [Oliveira et al. 2016]. Este método explora a ideia de que as sentenças que têm maior semelhança com o título têm maior relevância, e assim devem receber um maior escore de importância. Este método é calculado conforme apresentado na Equação 7. Além disso, computamos uma variação deste método usando a similaridade do cosseno entre o vetor de palavras (representadas por seus radicais) da frase com o título, desconsiderando as *stopwords*.

$$SimTitulo(s_i) = \frac{P_{s_i} \cap P_t}{P_{s_i} \cup P_t} \tag{7}$$

- P_t é o conjunto de palavras do título do documento que não são *stopwords*.

Bushy Path. Este método é baseado na ideia que frases importantes possuem muitas informações compartilhadas com outras sentenças [Oliveira et al. 2016]. Esta técnica computa a importância de uma sentença s, que é representada como um vértice em um grafo, calculando o total de arestas que a frase possui no grafo. Uma aresta entre duas sentenças é criada quando a similaridade do cosseno, ou qualquer outra medida de semelhança, entre essas duas frases for maior que um dado limiar (neste trabalho adotamos $0,1$ como limiar baseado em outros trabalhos da literatura). A ideia deste método é que sentenças que sejam altamente conectadas terão mais informações relevantes da coleção de documentos. A Equação 8 demonstra como este método é calculado.

$$BushyPath(s_i) = grau(s_i) \tag{8}$$

- $S\ grau(s_i)$ retorna o total de arestas que a sentença s_i possui no grafo.

Similaridade Agregada. Este método também utiliza uma representação em grafos para computar a centralidade das informações das sentenças [Oliveira et al. 2016]. Ele é similar ao Bushy Path, porém nele a pontuação de uma sentença é dada pelo somatório dos pesos das arestas de cada vértice (frase) no grafo. O peso de uma aresta a_{ij} é dado pela similaridade do cosseno entre as sentenças s_i e s_j se ela for maior que um dado limiar ou zero caso contrário. Utilizamos o limiar como $0,1$ baseado em outros trabalhos da literatura. O método é computado pela Equação 9.

$$SimAgregada = \sum_{i=1, j \neq j}^{S} peso_aresta(s_i, s_j) \tag{9}$$

TextRank. Este método baseado em grafos muito utilizado para a extração de palavras-chave em uma coleção de documentos [Oliveira et al. 2016]. Sua ideia é atribuir um maior escore para sentenças que possuem muitos n-gramas relevantes. A importância de um n-grama é determinada pela sua frequência e coocorrência com outros n-gramas frequentes. O escore de relevância de um n-grama é computado usando a Equação 10, enquanto que o escore de uma sentença baseado no TextRank é calculado como apresentado na Equação 11.

$$TextRank(v_i) = (1-d) + d \times \sum_{v_j \in In(v_i)} \frac{w_{ji}}{\sum_{v_k \in Out(v_j)} w_{jk}} \times TextRank(v_j) \tag{10}$$

$$TextRank(s_i) = \sum_{p_j \in s_i}^{P} TextRank(p_j) \tag{11}$$

- *d* é o fator de amortecimento que usualmente é definido com o valor de $0,85$;
- *In(v_i)* é o conjunto de vértices que apontam para v_i;
- *Out(v_j)* é o conjunto de vértices que v_j aponta;
- *P* é o total de palavras ou n-gramas da sentença s_i;
- p_j é uma palavra ou n-grama pertencente a sentença s_i;
- w_{ji} é o total de coocorrência entre v_i e v_j.

3. Experimentos

Nesta seção são discutidos os experimentos realizados para avaliar o desempenho dos métodos de ponderação das frases e dos algoritmos de regressão aplicados na etapa de mensuração da relevância das frases na tarefa de SAT multidocumento. Três experimentos foram executados, buscando apresentar respostas para a seguintes questões: (1) Quais das medidas de pontuação de frases avaliadas apresentam melhor desempenho? (2) Qual dos algoritmos de regressão obtém o melhor resultado na tarefa de estimar um escore de relevância de uma frase? e (3) Como os resultados obtidos estão se comparados com outros trabalhos na literatura?

3.1. Ambiente experimental

A avaliação dos resumos gerados por sistemas de SAT é uma tarefa desafiadora e de suma importância, pois estimula o desenvolvimento de novas abordagens, e assim o progresso da área de SAT [Lloret et al. 2017]. Um bom resumo, seja ele gerado automaticamente ou manualmente, deve ser informativo, ou seja, conter as informações mais relevantes dos textos originais, coeso, coerente e gramaticalmente correto.

Neste trabalho, dada a natureza extrativa da abordagem de SAT usada, focaremos somente na avaliação da informatividade dos resumos. Em geral, a avaliação da informatividade de um resumo na SAT é realizada através da comparação desse resumo com outros de referência criados manualmente. Neste trabalho, adotou-se o corpus CSTNews[Dias et al. 2014], criado para subsidiar pesquisas na área de SAT de artigos de notícias escritos em português do Brasil. O corpus CSTNews possui 50 grupos de textos, e em cada grupo existem aproximadamente 3 artigos jornalísticos de fontes diferentes abordando o mesmo assunto. Cada grupo de notícias possui 5 resumos extrativos criados para serem utilizados na avaliação de sistemas de sumarização multidocumento.

Escores de Relevância das Frases. O corpus CSTNews possui um conjunto de 5 resumos de referência extrativos para cada um dos seus grupos de documentos. Neste trabalho, para que sejam treinados os modelos de regressão é preciso ter uma pontuação de relevância das frases dos grupos de documentos, já que é essa pontuação será usada como atributo alvo. O corpus CSTNews não possui essa informação. Portanto, para gerar as pontuações de relevância das frases, foi utilizada a estratégia de calcular a similaridade do cosseno de cada frase do grupo de documento com cada frase inserida no seu respectivo conjunto de resumos de referências. Dessa forma, frases dos documentos originais que foram usadas para criar os resumos de referência possuem maior escore. Os valores dos escores foram normalizados e variam de 0 até 1. Para o calculo da similaridade do cosseno, cada frase foi representada por um vetor contendo suas palavras, sendo removidas as *stopwords* e as palavras sendo representadas por seus radicais.

Medida de Avaliação. Neste trabalho foi utilizado os indicadores de avaliação disponibilizadas pela ferramenta ROUGE, em especial foram adotadas as medidas de Cobertura, Precisão e Medida-F derivadas pela variação do ROUGE-1 [Lin 2004]. O ROUGE-1 computa a sobreposição de palavras entre os resumos gerados automaticamente e os resumos de referência. Cada grupo de documentos do corpus CSTNews possui mais de um resumo de referência. Por isso, os escores das medidas de avaliação adotadas nos experimentos são gerados, computando a média entre a comparação feita do resumo gerado automaticamente com os resumos de referência do corpus.

Tamanho resumos. Neste trabalho, os resumos gerados possuem aproximadamente 100 palavras. Esse limiar de tamanho do resumo foi adotado por ser comumente usado em diversos trabalhos de SAT na literatura. Para promover uma comparação justa, durante a avaliação usando o ROUGE, somente as 100 primeiras palavras dos resumos foram consideradas.

3.2. Avaliação dos métodos de ponderação das frases

Neste experimento são avaliados os métodos de ponderação das frases considerados neste trabalho. As medidas de Cobertura, Precisão e Medida-F derivadas da variação do ROUGE-1 (R-1) foram adotadas para avaliar a informatividade dos resumos gerados. Além dessas medidas, verificou-se a correlação de Pearson entre os escores gerados por cada medida em comparação com a pontuação de relevância de cada frase. A correlação de Pearson produz valores que variam entre -1 e 1, sendo quanto mais próxima de -1 isso representa uma correlação negativa entre as duas variáveis, ou seja, enquanto uma aumenta, a outra sempre diminui. Uma correlação mais próxima de 1 representa uma correlação positiva entre as duas variáveis, ou seja, quando uma aumenta a outra também aumenta. Quanto mais próxima de 0, isso significa que as duas variáveis não dependem linearmente uma da outra.

Na Tabela 1 são apresentados o coeficiente da correlação de Pearson (P), e as medidas de cobertura, precisão e a medida-F derivadas do ROUGE-1 em porcentagem para cada um dos métodos de ponderação analisados. Os melhores resultados em cada medida de avaliação são destacados em negrito.

Tabela 1. Resultados da avaliação e desvio padrão (entre parênteses) dos métodos de ponderação de frases.

Métodos de Pontuação	P	Cobertura (%)	Precisão (%)	Medida-F (%)
Bushy Path	0,2210	50,56 (11,68)	49,69 (9,64)	50,12
Centr. Frases Sim. Cosseno	0,2731	53,81 (12,11)	52,57 (9,76)	53,18
Centr. Frases Inters.	**0,3002**	59,41 (11,05)	56,94 (9,43)	58,15
Freq. Palavras	0,2440	51,46 (11,44)	49,92 (10,98)	50,67
FT-FIS	0,2672	54,11 (11,03)	52,53 (10,12)	53,31
Posiçao Frases	0,2666	**60,87** (9,08)	**59,83** (8,09)	**60,35**
Qtde Entidade Nomeadas	0,1460	51,66 (12,60)	51,14 (10,12)	51,40
Sim. Agregada	0,2689	52,70 (12,73)	51,68 (9,88)	52,19
Sim. Cosseno Título	0,2347	57,00 (10,67)	55,64 (9,03)	56,31
Sim. Título Inters.	0,2308	57,18 (10,31)	55,67 (9,25)	56,42
Textrank	0,1878	50,09 (11,10)	49,06 (10,79)	49,57

O método de Centralidade das Frases usando Intersecção (Centr. Frases Inters.) obteve a maior correlação de Pearson $(0,3002)$. Esse resultado demonstra que frases que compartilham mais informações com outras sentenças no grupo de documentos tendem a serem importantes e terem uma alta probabilidade de serem incluídas no resumo a ser gerado. Os valores de correlação obtidos pelos métodos de ponderação podem ser considerados, como correlações desprezíveis $(0,0 \leq |P| < 0,3)$. Esses valores são esperados devido à complexidade e subjetividade relacionada à tarefa de criação de resumos a partir de múltiplos documentos.

O método de posição das frases apresentou o melhor resultado nas medidas de avaliação de cobertura, precisão e medida-F. Tal resultado evidencia que frases que

estão mais próximas do início dos documentos possuem uma concentração maior de informações relevantes que devem ser inseridas nos resumos gerados. Esses resultados corroboram com outros trabalhos na literatura [Oliveira et al. 2016] que também observaram esse comportamento. Os métodos de centralidade das frases usando intersecção, similaridade do cosseno com o título (Sim. Cosseno Título) e sua variação usando intersecção das palavras também obtiveram bons resultados nas medidas de avaliação do ROUGE-1.

Apesar dos baixos valores obtidos na correlação de Pearson pelos métodos de poderação, eles apresentaram resultados razoáveis nas medidas de avaliação do ROUGE. Em especial, analisando a medida de cobertura é possível perceber que todos os métodos obtiveram resultados superiores a 50,00%, o que indica o percentual de informações que estão presentes nos resumos de referência que foram inseridos nos resumos gerados.

3.3. Avaliação dos algoritmos de regressão

O objetivo deste segundo experimento é avaliar o desempenho de diferentes algoritmos de regressão usados para estimar um escore de relevância das frases de uma coleção de documentos a ser resumido. Para isso, as medidas de ponderação das frases avaliadas na Seção 3.2 foram usadas como atributos para treinar os modelos de regressão. Neste experimento todos os métodos foram usados, ou seja, nenhum algoritmo de seleção de atributos foi adotado.

Os seguintes algoritmos disponíveis na ferramenta Scikit-learn[5] foram analisados: Regressão Bayesiana, Árvore de Decisão, Rede Elástica, Regressão de Huber, K-vizinhos mais próximos, do inglês *K-nearest neighbors* (KNN) com $k = 5$, Regressão Lasso, Regressão Linear, Regressão Ridge, Gradiente Descendente Estocástico, do inglês *Stochastic Gradient Descent* (SDG), Máquina de vetores de suporte usando como núcleo uma função de base radial, do inglês *Support Vector Regression with Radius Base Function* (SVR-RBF). Esses algoritmos foram escolhidos por serem tradicionalmente adotados na tarefa de regressão na literatura. Todos os algoritmos foram utilizados usando suas respectivas configurações padrões, ou seja, nenhuma calibração dos parâmetros foi realizada.

A metodologia de validação cruzada com k conjuntos (k-*fold cross validation*) foi adotada neste experimento para avaliar o desempenho dos dez algoritmos de regressão supracitados. O parâmetro k foi definido com o valor cinquenta ($k = 50$) porque o corpus CSTNews possui 50 grupos de documentos. Dessa forma, para cada grupo de documentos, o processo de geração do resumo é executado de acordo com as duas seguintes etapas:

1. **Treinamento:** Nesta etapa, os $k - 1$ grupos de documentos, ou seja, as frases de 49 grupos de documentos são usadas para treinar o modelo de regressão.
2. **Geração do resumo (Teste)**: O grupo de documentos não selecionado na etapa de treinamento é usado para teste. O modelo de regressão criado na etapa anterior é aplicado para estimar a medida de relevância de cada frase do grupo de documentos de teste e, posteriormente, é realizado o processo de sumarização dessa coleção seguindo o processo descrito na Seção 2.

Na Tabela 2 são apresentados os resultados deste experimento considerando as medidas de: **(i)** Correlação de Pearson (P); **(ii)** A média da soma dos quadrados residu-

[5]https://scikit-learn.org/stable/

ais, do inglês, *Residual Sum of Squares* (RSS) (ver Equação 12); e **(iii)** As medidas de cobertura, precisão e medida-F do ROUGE-1 ($R-1$) dos resumos gerados.

$$\frac{\sum_f^F (e_f - \overline{e}_f)^2}{|F|} \tag{12}$$

- F é o conjunto de frases existentes no corpus CSTNews;
- $|F|$ é o total de frases existentes no corpus CSTNews;
- e_f é o real escore de relevância da frase f;
- $\overline{e}_f$ é o escore de relevância da frase f estimado pelo algoritmo regressão.

Tabela 2. Resultados da avaliação e desvio padrão (entre parênteses) dos algoritmos de regressão.

Algoritmos	P	RSS	Cobertura (%)	Precisão (%)	Medida-F (%)
Árvore de Decisão	0,2601	0,1057	59,59 (10,15)	57,93 (8,54)	58,75
KNN	0,3240	0,0950	55,58 (9,67)	54,75 (8,02)	55,16
Rede Elástica	0,3364	0,0883	60,19 (9,79)	58,54 (8,40)	59,36
Regressão Bayesiana	0,4721	0,0790	**62,09** (9,69)	**60,12** (8,18)	**61,09**
Regressão Huber	0,4914	0,0758	61,22 (9,07)	59,59 (8,31)	60,39
Regressão Lasso	0,3359	0,0884	60,31 (9,75)	58,65 (8,36)	59,47
Regressão Linear	0,4914	0,0770	60,49 (9,06)	59,10 (7,95)	59,78
Regressão Ridge	0,4834	0,0781	61,17 (9,03)	59,55 (8,20)	60,35
SDG	0,3694	0,0860	59,90 (10,12)	57,87 (7,68)	58,87
SVR-RBF	**0,5608**	**0,0753**	58,13 (10,64)	57,08 (8,65)	57,60

Os algoritmos SVR-RBF, Regressão Linear e de Huber apresentaram as correlações de Pearson mais fortes com os valores $0,5608$ e $0,4914$, respectivamente. A correlação obtida pelo algoritmo SVR-RBF é considerada Moderada ($0,5 \leq |P| < 0,7$). Em geral, a combinação dos métodos de ponderação das frases em conjunto com os algoritmos de regressão investigados levou a resultados melhores em termos da medida de correlação de Pearson. Esse resultado demonstra que combinar essas medidas pode gerar escores estimados que seguem um comportamento mais próximo dos reais valores de relevância das frases. O algoritmo SVR-RBF ainda apresentou o menor valor de erro na medida de RSS, demonstrando que os escores estimados por esse algoritmo foram os que mais se aproximaram dos reais escores de relevância das frases.

O algoritmo de Regressão Bayesiana apresentou os melhores valores nas medidas de cobertura, precisão e medida-F do ROUGE-1. O resultado obtido na medida de cobertura demonstra que esse algoritmo conseguiu gerar resumos que possuem 62,09% das informações relevantes presentes nos resumos de referência. É importante destacar que o algoritmo de Regressão Bayesiana obteve 60,12% na medida de precisão e 61,09% na medida-F. Esses resultados indicam que os resumos gerados pelo algoritmo apresentam um bom equilíbrio entre a quantidade informações relevantes presentes nos resumos de referência e aquelas que o algoritmo selecionou para compor o resumo automático gerado.

3.4. Comparação com outros trabalhos

Neste último experimento é realizado uma comparação dos resultados obtidos pelo sistema desenvolvido usando o algoritmo de Regressão Bayesiana com outros três

sistemas identificados na literatura, sendo eles: GistSumm [Pardo 2005], CSTSumm [Castro Jorge and Pardo 2010], e RC-4 [Cardoso and Pardo 2016]. Na Tabela 3 são apresentados os resultados obtidos em termos das medidas do R-1. O melhor sistema em cada medida de avaliação é destacado em negrito.

Tabela 3. Resultados da comparação e desvio padrão (entre parênteses) com outros trabalhos da literatura.

Sistemas	Cobertura (%)	Precisão (%)	Medida-F (%)
CSTSumm	53,94 (11,42)	55,97 (9,10)	54,93
GistSumm	57,64 (11,21)	54,71 (9,31)	56,14
RC-4	59,10 (6,93)	**60,18** (9,24)	59,64
Sistema desenvolvido	**62,09** (9,69)	60,12 (8,18)	**61,09**

O sistema desenvolvido conseguiu obter melhores resultados do que os outros sistemas considerados nas medidas de cobertura e medida-F do ROUGE. Em especial, observa-se uma maior diferença em relação à medida de cobertura, o que indica que o sistema desenvolvido produziu resumos mais informativos do que os gerados pelos outros sistemas comparados. O sistema RC-4 obteve o melhor desempenho na medida de precisão e o segundo melhor resultado em relação as outras medidas. Esse sistema foi o que apresentou melhores resultados em comparação com os outros dois sistemas CSTSumm e GistSumm.

Os resultados experimentais demonstraram que a estratégia de usar algoritmos de regressão para combinar os métodos de ponderação das frases apresentou resultados competitivos com os outros sistemas da literatura considerados.

4. Considerações Finais e Trabalhos Futuros

Neste trabalho foi apresentado uma análise comparativa entre diferentes algoritmos de regressão para estimar a relevância de uma frase na tarefa de sumarização automática de artigos de notícias escritos em português do Brasil. Diversos métodos de ponderação da relevância das frases que são tradicionalmente usados na literatura na tarefa de SAT multidocumento foram investigados em conjunto com os algoritmos de regressão. Para realizar a análise comparativa, foi desenvolvido um sistema de sumarização multidocumento do tipo extrativo. Experimentos foram realizados usando o corpus CSTNews [Dias et al. 2014] adotando as tradicionais medidas de avaliação do ROUGE [Lin 2004].

Os resultados experimentais evidenciaram que os métodos de posição das frases, centralidade das frases usando intersecção e similaridade do cosseno com o título, apresentaram os três melhores resultados com base nas medidas de avaliação do ROUGE. A combinação dos métodos de ponderação das sentenças usando algoritmos de regressão melhorou os resultados obtidos com base nas medidas do ROUGE-1. O algoritmo de Regressão Bayesiana obteve os melhores resultados na comparação com outros algoritmos de regressão investigados com base nas medidas do R-1. O sistema criado apresentou resultados competitivos em comparação com outros trabalhos da literatura, obtendo o melhor desempenho nas medidas de cobertura e medida-F do R-1.

A principal contribuição deste trabalho é a investigação da aplicação de algoritmos regressão para a tarefa de sumarização multidocumento de artigos de notícias escritos em português do Brasil. Trabalhos na área de SAT em português do Brasil ainda são escassos,

se comparados com outros idiomas, como o inglês. Um dos pontos que limitam o trabalho foi a opção de desenvolver um sistema de natureza extrativa, o que limita o processo por não poder realizar nenhuma alteração nas frases que irão compor o resumo gerado.

A criação de resumos automaticamente é uma tarefa desafiadora e ainda existem muitos problemas de pesquisa a serem resolvidos. Diante disso, como futuros trabalhos pretendemos: (i) Incorporar mais métodos de ponderação de sentenças; (ii) Avaliar a aplicação de algoritmos de seleção de atributos para identificar os métodos de ponderação mais significativos; (iii) Analisar o impacto de ajustes nos hiperparâmetros dos algoritmos de regressão; e (iv) Analisar o uso de algoritmos de regressão baseados em redes neurais seguindo uma abordagem de aprendizado profundo (*Deep learning*);

Referências

Cardoso, P. C. and Pardo, T. A. (2016). Multi-document summarization using semantic discourse models. *Procesamiento del Lenguaje Natural*, (56):57–64.

Castro Jorge, M. L. d. R. and Pardo, T. A. S. (2010). Experiments with cst-based multidocument summarization. In *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs-5, pages 74–82, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dias, M. S., Garay, A. Y. B., Chuman, C., Barros, C. D., Maziero, E. G., Nobrega, F. A. A., Souza, J. W. C., Cabezudo, M. A. S., Delege, M., Jorge, M. L. R. C., Silva, N. L., Cardoso, P. C. F., Balage Filho, P. P., Condori, R. E. L., Marcasso, V., Felippo, A. d., Nunes, M. d. G. V., and Pardo, T. A. S. (2014). Enriquecendo o córpus cstnews: a criação de novos sumários multidocumento. In *International Conference on Computational Processing of the Portuguese Language - PROPOR*. SBC.

Leite, D. S. and Rino, L. H. (2008). Combining multiple features for automatic text summarization through machine learning. In *Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language*, PROPOR '08, pages 122–132, Berlin, Heidelberg. Springer-Verlag.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens, S. S., editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lloret, E., Plaza, L., and Aker, A. (2017). The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*.

Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 43–76. Springer.

Oliveira, H., Ferreira, R., Lima, R., Lins, R. D., Freitas, F., Riss, M., and Simske, S. J. (2016). Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Syst. Appl.*, 65(C):68–86.

Pardo, T. A. S. (2005). Gistsumm-gist summarizer: Extensões e novas funcionalidades. *Série de Relatórios do NILC*.